Université de Versailles Saint-Quentin-en-Yvelines

UFR des sciences

Département d'Informatique

Contrefaçon de logiciels dans le monde Evolution dans le temps et dans l'espace.

03 mai 2021

Encadré par

Stéphane LOPES

Réalisé par

Nadjib RAHMANI.

Toufik GUENANE.

Koussaila ARAB.

2020/2021

Table des matières

1 Introduction

La visualisation des données est l'une des parties les plus essentielles du pipeline de la science des données. Lors des premières étapes d'un projet de données massives. Souvent, une analyse exploratoire des données est effectuée afin d'avoir une compréhension plus profonde des données en question. Parfois, les données récoltées sont ambiguës tant qu'ils ne sont pas traités et mise sous une forme visuelle(Courbes et cartes ..... Etc)

La visualisation des données participe vraiment à rendre les choses plus accessibles et plus faciles à comprendre, en particulier avec des jeux de données de grandes dimensions et de grandes tailles. Au moment de finalisation du projet, il est important de pouvoir présenter des résultats finaux de manière simple, concise, convaincante et du qualité afin de donner au client une meilleure version compréhensible et du qualité des donnée étudier afin d'améliorer ses affaires.

L'enjeu majeur de ce projet est de réaliser les corrélations nécessaires à partir des données initiales traité et visualiser afin de donner la meilleure solution possible pour l'étude de client sur le marché de contrefaçon de logiciels dans le monde et son évolution dans le temps et dans l'espace.

2 Besoins du client

3 Choix du langage de programmation

3.1 Python

Python est un langage de programmation polyvalent doté d'un vaste ensemble de bibliothèques déjà existantes. Par conséquent, il peut facilement être utilisé pour développer des applications scientifiques et numériques qui requièrent toutes deux une grande complexité. Comme le langage est déjà conçu pour faciliter l'analyse et la visualisation des données, il est excellent pour les solutions personnalisées des données massives. Enfin, les bibliothèques et API de visualisation de données existantes permettent de visualiser et de présenter les données de manière attrayante et efficace.

3.2 Bibliothèques utilisées

Pandas

Comme les sources de nos données se présentent sous forme de feuilles Excel comportant un grand nombre de lignes et de colonnes. L'extraction et l'analyse de ces données nécessitent généralement une puissance de calcul élevée et sont considérées comme très longues. C'est pour cela que l'utilisation de la bibliothèque Pandas est devenue primordiale pour le projet, elle offre un traitement en parallèle de ces données ce qui accélère l'analyse et l'extraction de nos données.

Dash

Dash est un framework Python productif pour la création d'applications d'analyse Web. Écrit au-dessus de Flask, Plotly.js et React.js. Dash est idéal pour créer des applications de visualisation de données avec des interfaces utilisateur hautement personnalisées en Python pur. Il est particulièrement adapté à quiconque travaille avec des données en Python.

Tkinter

Tkinter est la bibliothèque GUI standard pour Python. Python lorsqu'il est combiné avec Tkinter, fournis un moyen rapide et efficace de créer des applications GUI. Tkinter fournit une puissante interface orientée objet à la boîte à outils Tk GUI et celà nous permet de personnalisé le travail de visualisation d'une façon concise et efficace.

Matplotlib

Matpoltlib est le package Python le plus utilisé pour les graphiques 2D. Il fournit à la fois un moyen rapide de visualiser les données de Python et des chiffres de qualité publication dans de nombreux formats. Matplotlib est livré avec un ensemble de paramètres par défaut qui permettent de personnaliser toutes sortes de propriétés. Cela nous permet de contrôler les valeurs par défaut de presque toutes les propriétés de matplotlib : taille de la figure et dpi, largeur de ligne, couleur et style, axes, propriétés de l'axe et de la grille, propriétés du texte et de la police. Ces qualités font de matplotlib une bibliothèque très intéressante et puissante qui répond à tous nos besoins.

Plotly

Après l'analyse des données, elles doivent être présentées dans un format facilement compréhensible. C'est là où la visualisation des données entre en jeu. Plotly nous permet de créer des visualisations interactives sur un navigateur web tel que de diagrammes de dispersion, des histogrammes, des diagrammes bâton et des cartes géographiques.

psycopg2

Psycopg est l'adaptateur de base de données PostgreSQL le plus populaire pour le langage de programmation Python. L'adaptation peut être étendue et personnalisée grâce à un système flexible d'adaptation d'objets.

4 Création et gestion de bases de données

4.1 SGBD : PostgreSQL

PostgreSQL est un système de gestion de bases de données relationnelles et objet (SGBDRO). Il supporte une grande partie du standard SQL tout en offrant de nombreuses fonctionnalités modernes comme : requêtes complexes, vues modifiables et intégrité transactionnelle. De plus, PostgreSQL peut être étendu par l'utilisateur de multiples façons, en ajoutant : de nouveaux types de données, de nouvelles fonctions et de nouvelles fonctions d'agrégat, et l'utilisation de ce SGBD est primordial. [1]

4.2 Choix de données

Dans cette section nous allons listées et justifier les différents données utilisé pour la constitution dans notre base de données

Données Économique

Le PIB reste l'indicateur le plus utilisé pour illustrer la croissance économique et peut être utile pour comparer les performances économiques de différents pays.

On a décidé de choisir les données GDP per capita (PIB par habitants) qui tient en compte le «coût de vie » de chaque pays .

La GDP per capita mesure le pouvoir d'achat d'une monnaie pour un consommateur pour se procurer le même panier de biens et de services qu'un autre consommateur dans un autre pays. Contrairement au taux de change, ce taux de conversion entre les monnaies tient alors compte du « coût de la vie ». Il est donc plus près de la richesse réelle par habitant.
Ces données seront prochainement visualisé , analyser et comparer aux taux de piratage de chaque pays , dont les but est de savoir s'il existe une corrélation entre ces deux données différentes . [2]

Serveur sécurise

Le but d'un serveur sécurise, c'est de Protéger les données, Renforcer les mesures de sécurité appliquées aux logiciels et aux entreprises et centraliser un grand nombre de données alors nous avons choisi ces donnes pour savoir la réussite des serveurs sécurisé à diminuer le taux de piratage . [3]

Immigration

Ces données représentent le nombre d'immigrants vers un pays et le taux d'immigration par rapport à la population d'un pays, le but de ces donnes est de pouvoir étudier la corrélation avec le taux de piratage de logiciels. [4]

Régimes politiques des pays

Nous avons décidé d'intégrer ces donnés dans notre base de données afins de comparer la nature du régime politique des pays (Autocratie , Anocratie , démocratie ...) avec le taux de piratages de logiciel afin de voir s'il existe une corrélation entre ces deux données (Ex : Est-ce que le taux de piratage de logiciel est plus élevé dans les pays autocratique que dans les pays démocratique ?).[5]

Taux de chômage des pays

L'idée d'ajouter ces données a notre base de données et de pouvoir visualiser si l'augmentation ou la diminution des taux de chômage dans un pays choisi affectent l'augmentation ou la diminution du taux de piratage de logiciel dans ce pays.[6]

Utilisation Internet

Ces données représente le pourcentage de la population de chaque pays utilisant Internet, ils seront comparé aux de piratage de logiciel à fin d'en déduire si ces données favorise ou pas l'évolution du piratage de logiciels .[7]

Accès électricité

Ces données représente le pourcentage d'accès à l'électricité dans chaque pays , ils seront visualiser et comparé aux taux de piratage de logiciel dans le but est de savoir si ces deux données sont corréler .[8]

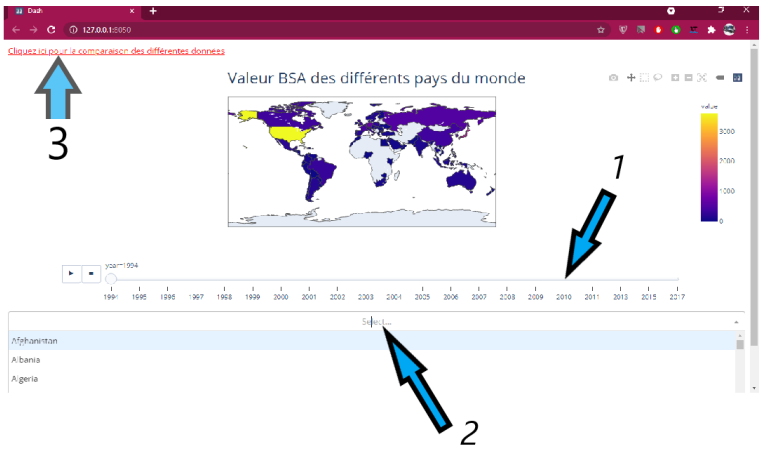
4.3 Connexion avec Excel

La connexion d'un fichier Microsoft Excel a notre base de données PostgreSQL se fait par l'installation d'un connecteur ODBC(Open Database Connectivity) .ce pilote a pour but d'importer les données directement dans une feuille de calcul Excel et les présenter sous forme de tableau . [9]

5 Exemple de manipulation

5.1 Programme

Une fois le programme est exécuté l'utilisateur se retrouve devant l'interface graphique suivante : L'utilisateur aura les choix suivants :

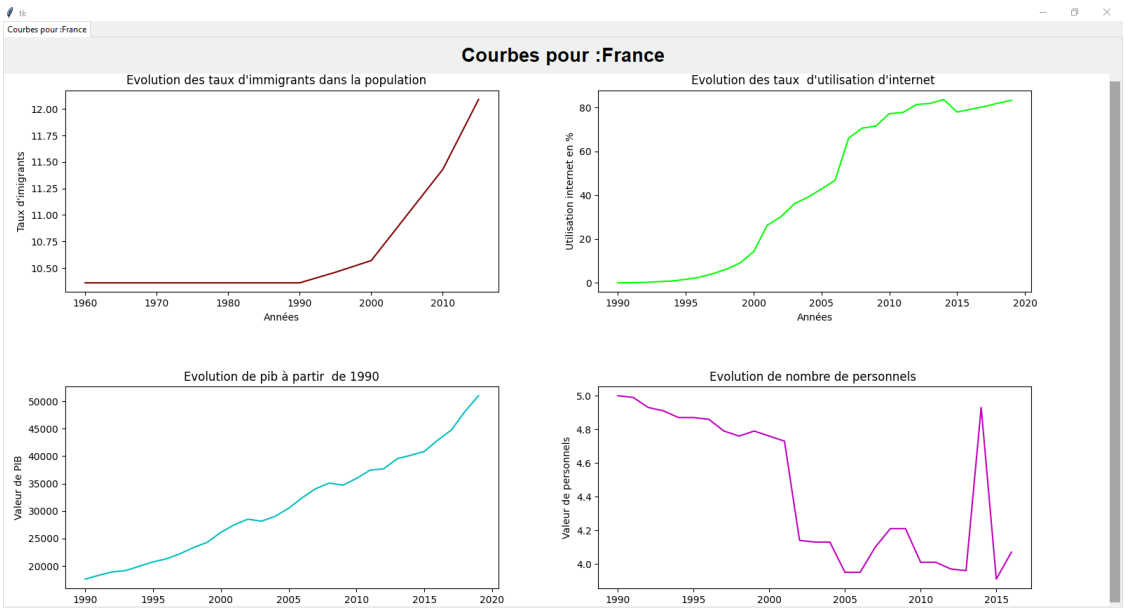


1

Visualiser sur la map les valeurs BSA de chaque pays en fonction d'une année choisie sur l'échelle représenté en dessus de la map .

2

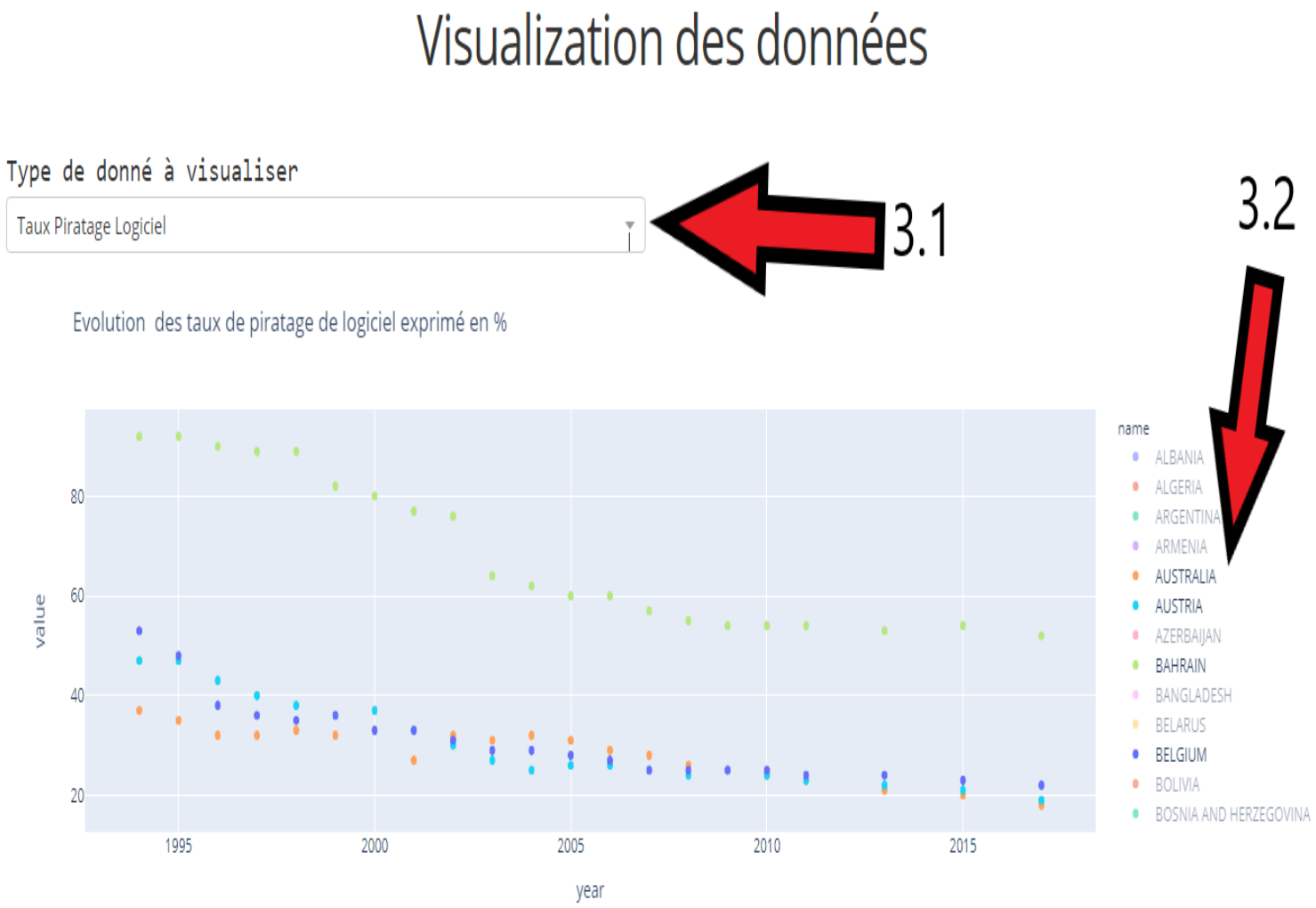
Sélectionner le pays qu'il souhaite afin de visualiser graphiquement les différentes données (taux de piratage, Pib, ...). Un exemple de la visualisation est représenté ci-dessus :



3

Visualiser et comparer graphiquement une donnée choisie entre les différents pays . Un exemple est représenté ci-dessus :

Cliquez ici pour la comparaison des différentes données

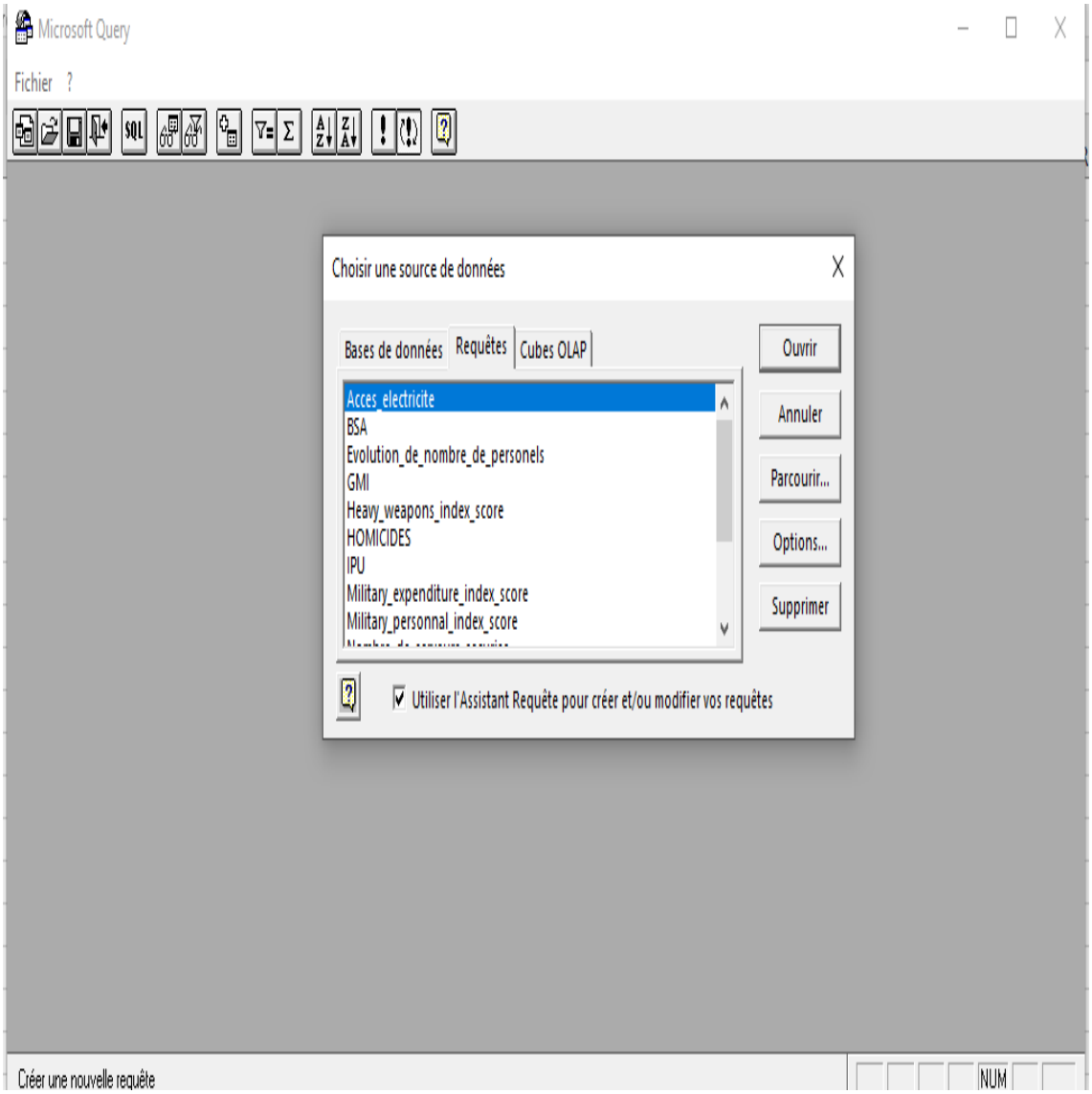


5.2 Excel

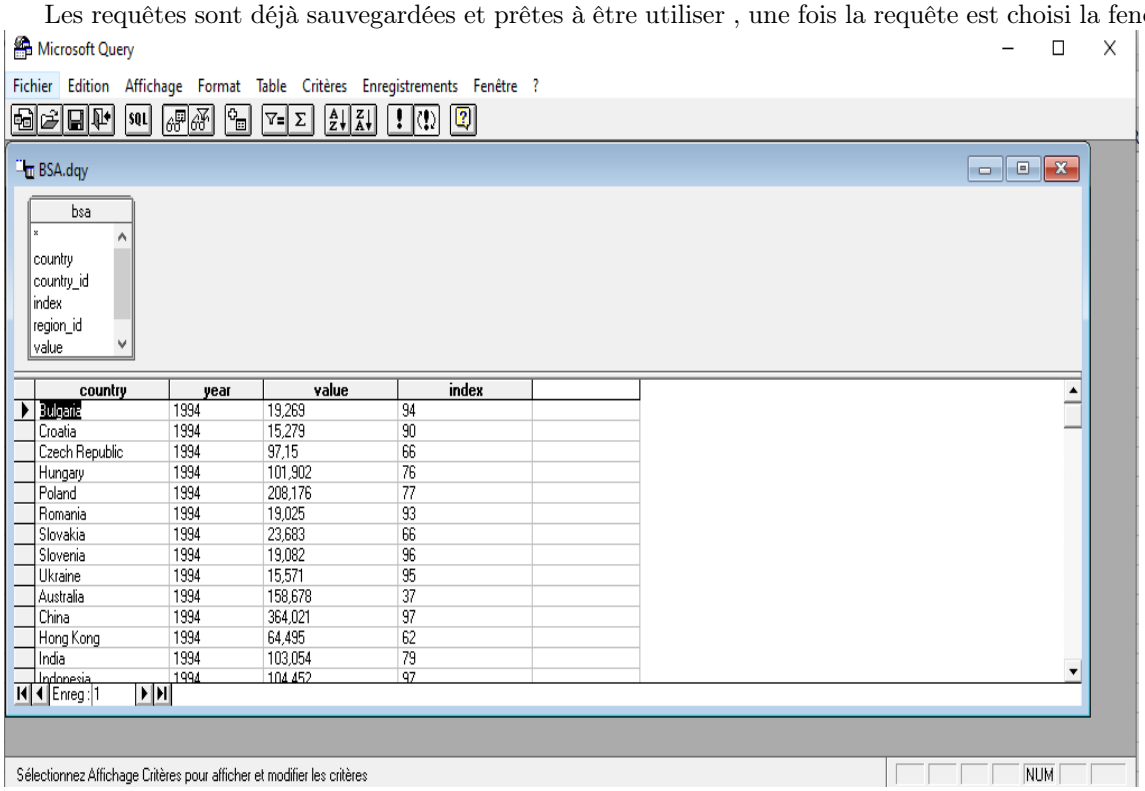
Pour visualiser les données sur un fichier Excel il faut choisir la section suivante :

Données -> données externe ->Microsoft query

La fenêtre suivante sera ouverte :



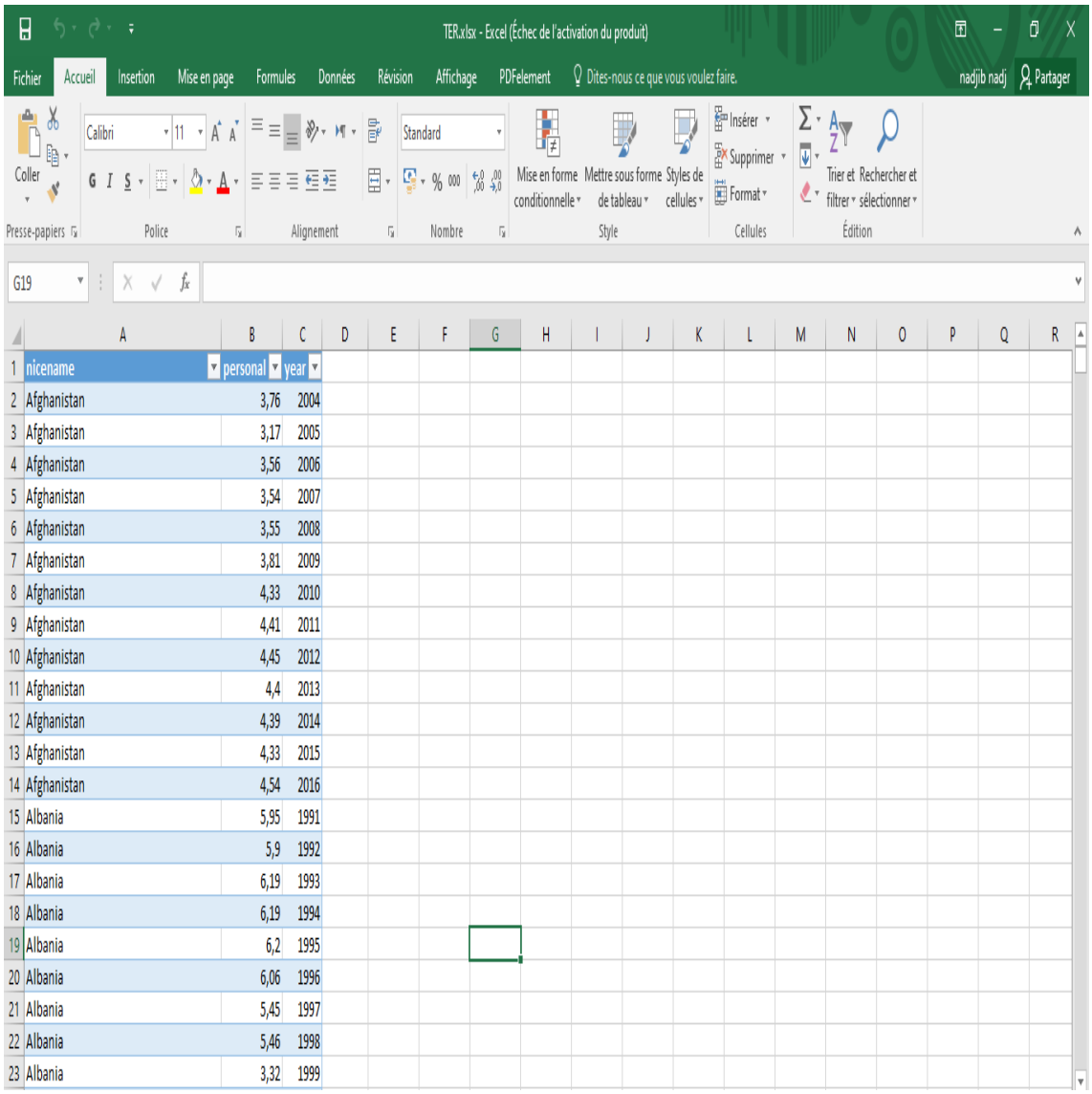
Les requêtes sont déjà sauvegardées et prêtes à être utiliser , une fois la requête est choisi la fenêtre suivante sera affiche :



4

filtre pour choisir un pays

En suite les données seront affichés sur Excel de la manier suivante :



	A	B	C
1	nicename	personal	year
2	Afghanistan	3,76	2004
3	Afghanistan	3,17	2005
4	Afghanistan	3,56	2006
5	Afghanistan	3,54	2007
6	Afghanistan	3,55	2008
7	Afghanistan	3,81	2009
8	Afghanistan	4,33	2010
9	Afghanistan	4,41	2011
10	Afghanistan	4,45	2012
11	Afghanistan	4,4	2013
12	Afghanistan	4,39	2014
13	Afghanistan	4,33	2015
14	Afghanistan	4,54	2016
15	Albania	5,95	1991
16	Albania	5,9	1992
17	Albania	6,19	1993
18	Albania	6,19	1994
19	Albania	6,2	1995
20	Albania	6,06	1996
21	Albania	5,45	1997
22	Albania	5,46	1998
23	Albania	3,32	1999

6 Conclusion

Bibliographie

[1] Postgresql. https://www.postgresql.org/.

[2] Données Économique source. https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD.

[3] Serveur sécurise source. https://data.worldbank.org/indicator/IT.NET.SECR.P6//.

[4] Immigration source. https://www.postgresql.org/ftp/odbc//.

[5] Régimes politiques source. https://ourworldindata.org/democracy#world-maps-of-political-regimes-over-200-years.

[6] Taux de chômage source. https://www.macrotrends.net/countries/ranking/unemployment-rate.

[7] Utilisation internet source. https://data.worldbank.org/indicator/IT.NET.USER.ZS.

[8] acceés électricité. https://www.macrotrends.net/countries/ranking/electricity-access-statistics.

[9] Odbc drvier. https://www.postgresql.org/ftp/odbc//.